Assessment System Design Options for the Next Generation Science Standards (NGSS) Reflections on Some Possible Design Approaches

Kathleen Scalise

April 2014

K-12 Center at ETS

Invitational Research Symposium on Science Assessment

Assessment System Design Options for the Next Generation Science Standards (NGSS): Reflections on Some Possible Design Approaches

Kathleen Scalise

University of Oregon

Introduction

The purpose of this paper is to provide an analysis of possible high-level designs of comprehensive science assessment systems that might be suitable to the U.S. Next Generation Science Standards (NGSS) context. In this paper, I first describe some overall ways that system design can be approached, drawing on the National Research Council (NRC) assessment triangle and concepts of evidence-centered design practices, applied here to system design. Then I offer three possible high-level designs of comprehensive science assessment systems. These are purposively sampled to span from the less familiar in the U.S. context to the more familiar, to provide a range of contrast and more clearly show how design decision-making frameworks might be employed. For additional contrast, I also include in this paper a brief “nonexample” of a system design that would likely have more limited utility for the NGSS context, based on evidence-centered thinking around the goals of the student model. Finally, I close with a brief recap and summary of such design approaches, as well as some recommendations for use.

Why a System for Science Assessment

The NRC has released two reports in recent years describing why systems, or multiple opportunities for obtaining assessment information, are needed for science education in the United States. The first report, *Systems for State Science Assessment* (National Research Council [NRC], 2005), examined K-12 science assessment in the United States. The report explored the concept that to assess science learning at the state level generally requires a system, or set of coordinated processes and instruments, for effective assessment.

In this paper, systems are defined as including a variety of components and strategies for collecting an appropriate array of assessment information for the multiple intended purposes, which may include both summative and formative aspects. The NRC (2005) report described systems of educational assessment as intended to answer a range of questions and serve information needs at different degrees of specificity for educational uses.

Building on this understanding, the NRC recently released a prepublication version of the 2014 report, *Developing Assessments for the Next Generation Science Standards*, with final report expected soon. The new report also described a system focus as needed for NGSS assessment. For instance, the report pointed out that any effective system of science assessment needs to include assessments both grounded in the classroom and assessments that provide information about the effectiveness of instruction and the overall progress of students' science learning.

The 2014 NRC report also expressed that states need to tailor their plans to their own circumstances and needs, while at the same time helping teachers, students, and schools acquire the information they need to stay on track with providing and obtaining the opportunity to learn appropriately in science education. As the report described, these are important goals of assessment—but also demanding goals. So satisfying them all with a strong degree of appropriateness in the data collection, or observations, and accumulation of evidence to make inferences about learning likely requires a range of data, through a systems approach, according to the NRC (2014).

For science education and the NGSS in particular, the challenges for assessment are several. Other commissioned papers for the Invitational Research Symposium on Science Assessment have pointed out that, while assessment tasks generally for all of education must elicit evidence related to student learning, these tasks must do more than this for the NGSS (DeBarger, Penuel, & Harris, 2013). The NGSS assessments must elicit evidence related to students' integration of knowledge of disciplinary core ideas, engagement with scientific practices, and facility with building connections across ideas (Pellegrino, 2013). In some aspects, these are hard-to-measure constructs (Haertel et al., 2012; Scalise, 2012). The blending or fusing of the student performance in this way as described by the NGSS framework and standards is wonderfully supportive of the educational experience in science; it is also challenging for assessment and measurement of learning outcomes.

Additionally, the NGSS exemplifies but often does not proscribe the full range of applications in which students may be instructed to meet the science performance expectations, or statements about what students should know and be able to do at each grade level. The NGSS framework developers have agreed that it is not possible to teach all possible applications of the NGSS, so determining which groups or "bundles" of performance expectations to assess also becomes a challenge (DeBarger et al., 2013; Quinn, Schweingruber, & Keller, 2012). For assessment, decisions must be made about whether and how to integrate disciplinary core ideas, science practices, and crosscutting concepts within tasks, instruments, rubrics, scales, and reports, as well as throughout all system-level components.

For all of these reasons, the NRC NGSS report (NRC, 2014) concluded that it will not be feasible to assess all of the performance expectations for a given grade level during a single assessment occasion. According to the report, students need multiple—and varied—assessment opportunities to demonstrate their competence on the performance expectations for a given grade level.

Additionally, the report (NRC, 2014) described other assessment information important to collect, such as the inputs to learning and an audit of the various aspects of opportunity to learn the NGSS both for student and teacher. Thus, the report called for a system of assessment. I next consider some principled ways by which to consider how to design such a system that might yield coherent and effective results.

NRC Assessment Triangle and Evidence-Centered Approaches

The approach illustrated here for system design is grounded in the NRC assessment triangle shown in Figure 1, which is an adaptation of the NRC (2001) assessment triangle for educational measurement.

The central challenge in assessment, according to the NRC (2001) report, is making inferences about attributes (the *cognition* vertex in Figure 1) that are not directly observable using a limited set of structured assessments (the *observations* vertex in Figure 1). The role of psychometric and statistical tools as well as other tools of descriptive information gathering and/or informal assessment is to negotiate legitimately the path from observations to inferred person measures (the *interpretation* vertex in Figure 1), often consistent with a theory of person change or identifying differences between persons. The assessment triangle is conceptualized as a process that is repeated multiple times in both assessment development and in the application of assessments to educational needs.

To date, the assessment triangle and associated elaborations mostly have been applied to the more fine-grained development of assessment tasks, or assessment instruments. Here I begin to apply the concept of the assessment triangle to design considerations of the overarching system characteristics themselves.

Similar and more elaborated approaches to the assessment triangle are described by Robert Mislevy's evidence-centered design (Mislevy, Almond, & Lukas, 2003; Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003), Mark Wilson’s BEAR Assessment System (Wilson, 2005), measuring learning progressions in science (Wilson, 2009), Embretson’s cognitive design system (Embretson & Reise, 2000), and other key approaches that have been established to meet the needs of the assessment circumstances (Hambleton, Swaminathan, & Rogers, 1991). To provide background, such developments emerged from earlier work spearheaded by Messick and others, where thinking regarding performance assessments with a construct-centered approach was based on asking what complex of knowledge, skills, or other attributes should be assessed, and next, what behaviors or performances would reveal those constructs, or elicit the behaviors (Messick, 1994). Messick described the importance of the selection or construction of relevant tasks and rational development of construct-centered scoring criteria and rubrics.

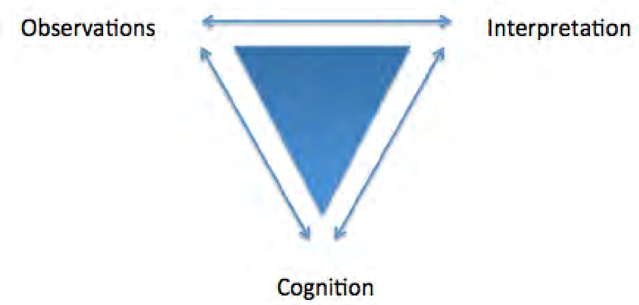


Figure 1. A depiction of concepts contained in the NRC assessment triangle.

Many of these evidence-based elaborations share important traits—and have some key differences in perspective. From the point of view of this paper, the key in the use for system design is that the development of the design is coherent among all its parts. The design needs to be based on a full evidentiary argument that supports the intended inferences and conclusions. In that way, the system design is *evidence-centered*, in the sense of the overarching school of thought that the variety of elaborations represent, rather than one particular instantiation, or formal model.

The takeaway should be that different systems and purposes will require different ways of thinking through the logic chain connecting the system elements—but there should be an explicit logic chain and a clear argument for how the pieces fit. This is also consistent with scientific thinking. Next I discuss traversing the assessment triangle, considering each component and its purpose, from an evidence-centered perspective. Three concepts will be addressed, each with some associated essential questions to ask in system design in upcoming sections. The three concepts are:

1. The conception of the cognition vertex, or student model for system design.
2. The conception of the observation vertex, or task model for system design.
3. The conception of the interpretation vertex, or the evidence model for system design.

Cognition: The Student Model

A systems-level design approach to the cognition vertex of the assessment triangle needs to ask questions about the goals and objects of education that are being measured. The cognition vertex for a *system* is exemplified by its theory of action by which the system is intended to impact student learning outcomes regarding these goals and objectives. Thus, for coherent system design for NGSS assessment, an explicit *theory of action* for *the assessments* is necessary for the student model.

Here, for this component, the essential questions to ask include:

- What is the system's theory of action for student learning impacts resulting *from the assessments* and the collection of evidence they allow?
- In what way will putting the designed system in place support student learning outcomes on the intended goals and objectives?
- Furthermore, what are the overarching components of the system that would need to be in place to support the learning outcomes, given the theory of action and the intended purposes of the assessments?

This is key for an adequate student model in evidence-centered design: As the name suggests, the design must be focused on the student. So a theory of action here is defined as the rationale behind the strategy. To specify a student model, one must ask what the rationale is by which the gathering of NGSS assessment information is actually intended to impact student learning outcomes.

A variety of rationales, or theories of action, might be conceived by which a state or locality might believe NGSS assessment information would prove helpful in student learning outcomes. So different student models might be described by different groups. For the purpose of illustrating system design, here I select a theory of action summarized in the NRC report, *Systems for State Science Assessment* (2005). The report described student learning outcomes improved in science education by assessment evidence that can guide instructional decisions, hold schools accountable for meeting learning goals, monitor program effectiveness, and signify and exemplify through its tasks the goals for student learning.

This theory of action is only one constellation of ideas that a state or locality might adopt as the rationale behind its science assessments. Whatever the rationale is, however, the key for system design is that the theory of action should be able to show how the rationale is supported by the evidence produced by the system. In other words, there should not be a disconnect in the system between what it purports to do and the data it collects and supplies. The theory of action should connect with the information the system is designed to collect.

Some specific challenges for assessing the student model of the NGSS are briefly outlined in the prior section. More in-depth discussion is outside the scope of this paper as the topic is the focus of other papers commissioned by the K-12 Center at ETS for the Invitational Research Symposium on Science including DeBarger et al. (2013). Readers, however, should keep in mind when considering the system design for the student model that NGSS assessments must elicit evidence related to students' integration of knowledge of disciplinary core ideas, engagement with scientific practices, and facility with building connections across ideas (Pellegrino, 2013). Additionally, the NGSS exemplifies but often does not proscribe the full range of applications in which students may be instructed. Therefore everything from task design and test assembly, to scoring and rubrics, to scales and reports, is influenced by these factors of the student model.

Observations: The Task Model

A system design approach to the observation vertex of the assessment triangle helps states and localities understand what types of tasks or other observations should be put in place in an assessment system. This aspect of the triangle asks questions of policymakers, state assessment leads, system developers, teachers, students, and other stakeholders about what they should be doing in the system, such as:

- What types of tasks must students achieve for the theory-of-action impacts of the assessment system to fall into place?
- Will the system help the intended outcomes to be supported and persist in student learning through these observations? If so, how? If not, why not, and does the system design need to be improved to better elicit the most helpful observations for the NGSS or other science assessment's purpose?
- Are there other observations besides tasks that would be helpful to collect, and if so, what are they?

The main messages of Pellegrino (2013) regarding NGSS tasks and observations as presented to the National Science Foundation are:

- The assessment tasks should allow students to engage in science practices in the context of disciplinary core ideas and crosscutting concepts.
- Multicomponent tasks that make use of a variety of response formats will be best suited for this.
- Selected-response questions, short and extended constructed-response questions, and performance tasks can all be used, but they should be carefully designed to ensure that they measure the intended construct and support the intended inference.
- Finally, students will need multiple and varied assessment opportunities to demonstrate their proficiencies with the NGSS performance expectations.

The student model here specifies that tasks should do a good job of signifying goals of the NGSS, which has large implications for the overarching task model of a systems design.

Regarding signifying goals, this means that the tasks of the assessment system itself are called to be a role model of what students should know and do and teachers to teach. In the case of this example, what students should know and do is defined by the NGSS.. The tasks, therefore, should not be primarily proxy indicators; they should be the real thing, to the extent possible. So what is the real thing for the NGSS?

Chapter 2 of the 2014 NRC report discussed types of tasks suited to assessing the NGSS. The task examples are not repeated here, but they illustrate tasks in which students are asked to apply scientific and engineering practices in the context of disciplinary core ideas and crosscutting concepts, while maintaining connections across concepts and disciplines.

Many of the task examples exemplify use, either in the classroom or for providing evidence for monitoring purposes and program evaluation, or both. Decontextualized questions that are not connected to any scientific contextual materials or activities before or after or that are examples of rote memorizations of declarative knowledge that do not blend the three NGSS objectives are discouraged.

While the NGSS does offer some unique challenges for observations as described previously, the overall premise of moving to deeper, richer, more contextualized assessment tasks and observations is not new in U.S. state thinking. For instance, a systems approach also recently was recommended in connection with assessing the Common Core State Standards in *Criteria for High-Quality Assessment* (Darling-Hammond et al., 2013).

Darling-Hammond et al. (2013) discussed how states must not only evaluate how students are doing on the standards but also whether students have had the opportunity to learn. So-called fill-in-the-bubble tests will not suffice. According to Darling-Hammond et al., other nations already employ many new types of performance assessments to assess how well students evaluate and use information rather than just testing for recall. The authors described how these assessments frequently ask students to demonstrate what they know through written, oral, mathematical, physical, and multimedia products, as the U.S. state consortia and other groups are increasingly doing in recent years.

In this case and for these kinds of assessments, the definition of observations, or *observables*, will be defined here as questions, actions, or processes by which one can observe the targeted knowledge, skills, and abilities. Potential observables in technology-based or hands-on science tasks could, for instance, include process data about how a student approached and completed an activity, or the reasoning facets they exhibited that they used to explore their ideas, as well as questioning formats such as described above.

In other words, an observable may go beyond a simple written question or even a group of questions and become some other type of interaction or elicitation that produces evidence about learning (Scalise, 2014). However, the relationship between the observable and the inference made from it should be clear, explicit, and transparent, not only for high-quality characteristics of the evidence but also so that educators can interpret the evidence. Once again, the coherence of the assessment design is key in making valid and reliable inferences.

Thus, a potential observable becomes an actual observable in the measurement schema when it is clearly mapped and validated to indicate how the information will add evidence to the interpretation and support the student model.

Interpretation: The Evidence Model

Having appropriate observations in a system of science assessment is an excellent way to support the *signifying* aspect of the theory of action. Observations alone, however, are not enough. As the NRC assessment triangle describes, there must be an evidence model, or a way to interpret what the observations are telling us about the student, school, program, or other aspect of the system that the design is intended to inform. In other words, how is one to interpret the performance or the observations?

A system design approach to the interpretation vertex of the assessment triangle maps the observations back to making an interpretation on the construct. This approach asks:

- What evidence should the system as a whole provide to achieve the theory of action? How will the observations be scored or interpreted? What is the sort of valuing or categorization possible to assign to the observations, to indicate which are examples of the more mastery states of learning and what are examples of the more emergent or early behavior on the construct?
- What inferences can or should be made and how can they be reported and used for good utility? What maps back to support the theory of action and establishes this utility in the system design?
- What in the system can support premises laid out in the construct and observation vortices of the triangle? Which stakeholders in the system need what kind of information, and how should it be supplied?

Three Examples of System Design

To illustrate the evidence model, in the next section I take up three possible system models of many potential examples and work through each system's view of the student model, the task model, and the evidence model.

The three designs selected to sample purposively illustrate some possible contrasts in design. Of course, actual systems that any state or locality might adopt would tend to blend components in less extreme ways. Indeed, many of the components here could be exchanged or swapped between the examples, given the caveat that a coherent argument can be made for how the components would work together and what they would achieve to fulfill the student model of improved learning outcomes through the use or availability of the assessment information. Actual systems would, of course, also need considerable more elaboration than shown here. However, thinking through how to make a coherent design argument early in the process is argued here as a useful and essential part of system design, which can be the takeaway from this illustration section.

The three examples explored in detail in the following sections in this paper are, as named here, (a) a curriculum-based model, (b) a common tasks model, and (c) a traditional with inverted emphasis model.

The three designs are described in detail in the following sections. As an overview, the three designs are exemplified in Figures 2–4, and the key aspects of each design are summarized in Table 1.

These three design models were first presented at the Invitational Research Symposium on Science Assessment, held in Washington, DC in September 2013, in the session “NGSS Assessment System Designs: What Are the Challenges, Choices, and Trade-Offs?” (Scalise, 2013a). Figures 2–4 illustrate models using similar icons from previous summaries of the assessment consortia system components for Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (Smarter Balanced), the two alternate assessment consortia (Dynamic Learning Maps [DLM] and National Center and State Collaborative [NCSC]), and the two English proficiency consortia (Assessment Services Supporting ELs through Technology Systems [ASSETS] and English Language Proficiency Assessment for the 21st Century [ELPA21).

Table 1. High-Level Summary of System Components for Three Design Options

Student model theory of action components	Example 1: Curriculum-based model	Example 2: Common tasks model	Example 3: Traditional with inverted emphasis model
Monitor	Accreditation model	Common tasks, input survey, NAEP	Matrix sample TEA
Instruction	State/locality adoption process	Shared extended common tasks	Classroom-based TEA
Program evaluation	Course completion data profiles	Secure extended common tasks	Classroom-based TEA, professional learning communities, data walls
Signifying	Tasks exemplify NGSS	Tasks exemplify NGSS	Tasks exemplify NGSS

Note. NAEP = National Assessment of Educational Progress, NGSS = Next Generation Science Standards, TEA = technology-enhanced assessment.

Curriculum-Based Model

In the curriculum-based model, the emphasis is on states and localities working at least in part through their usual adoption processes for ensuring effective materials are used in the classroom. Through the adoption process, most states or localities have a way already in place by which they help schools adopt educational materials designed in accordance with a variety of specifications. These may include books, study guides, online homework, assessments, web sites, teacher editions, and more.

In the curriculum-based model, states and localities help schools to assess the NGSS by specifically requesting that science education materials include in their development and certification the assessments that are appropriately aligned to the NGSS.

Specifications for high-quality curricula could include NGSS-aligned embedded assessments, performance tasks, professional development modules, and even perhaps end-of-unit or end-of-year assessments. Any of these might be technology-enhanced or not and draw on formats or approaches recommended in NRC (2014).

Embedded tasks might be scored by teachers, if desired, or through external moderation, automated scoring, and other approaches.

The key for the curriculum-based model is that local education agencies select curriculum as usual from the approved menu, and state monitoring might consist of course completion rates and success levels, such as who is completing (are all students meeting the learning targets), what are they completing, is the diversity of the classroom well supported, and are there hot spots of students lacking opportunity to learn through such courses?

Of course, states and localities would need to exercise their usual processes of due diligence in specifying and inspecting materials for adoption. Most adoption processes already have in place committees that engage in reviews of such materials on a rotating basis and mechanisms by which schools may make their selections, including for assessments. Generally these mechanisms are not yet in place for NGSS adoptions, being that states and localities only just are getting the opportunity to consider if, how, and in what ways they might work with NGSS ideas and concepts.

To take the model one step further, the curriculum-based model conceived as a system might additionally include some types of course certification, as takes place currently for Advanced Placement® (AP®) courses and International Baccalaureate (IB) courses in science education.

Should the adoption process support it, these course approaches or provision of course materials and assessments need not be reserved only for advanced or honors high school courses. In some communities, for instance, adopted access to such science courses with embedded assessments are already available for student course credit in the K-12 system, for instance through the Stanford University Education Program for Gifted Youth. This program has for many years provided distinguished science courses with elaborate assessment systems for children and adolescents of many ages, including in physics, mathematics, and computer programming.

Through Stanford and a number of other such programs, schools may already engage in such approaches, which at Stanford, for instance, already currently are aligned to the Common Core State Standards and might conceivably be aligned to the NGSS in the science areas if schools and localities showed interest in such approaches in their adoption models.

For monitoring purposes, course offerings and course completion of science courses using such materials, in schools involved in such adoptions under a curriculum-based model, could be reviewed by participating states and localities. Course completion data profiles could be used to identify schools that might indicate the need to bolster opportunity to learn for these students, including in their programs and outcomes.

Figure 2 shows a green curriculum, a red curriculum, and a purple curriculum. These colors are simply intended to indicate that different sets of materials could apply in the adoption process, as is usual in many states and localities. Figure 2 further indicates that depending on specifications that states or localities might choose to make, adoption-specified embedded assessment activities or processes might be included in the adoptions, performance tasks, and/or a variety of examinations. These might include periodic assessments dispersed at key intervals as in some of the Stanford courses, or end-of-course exams as in AP and IB, or some combination of any or all of the above.

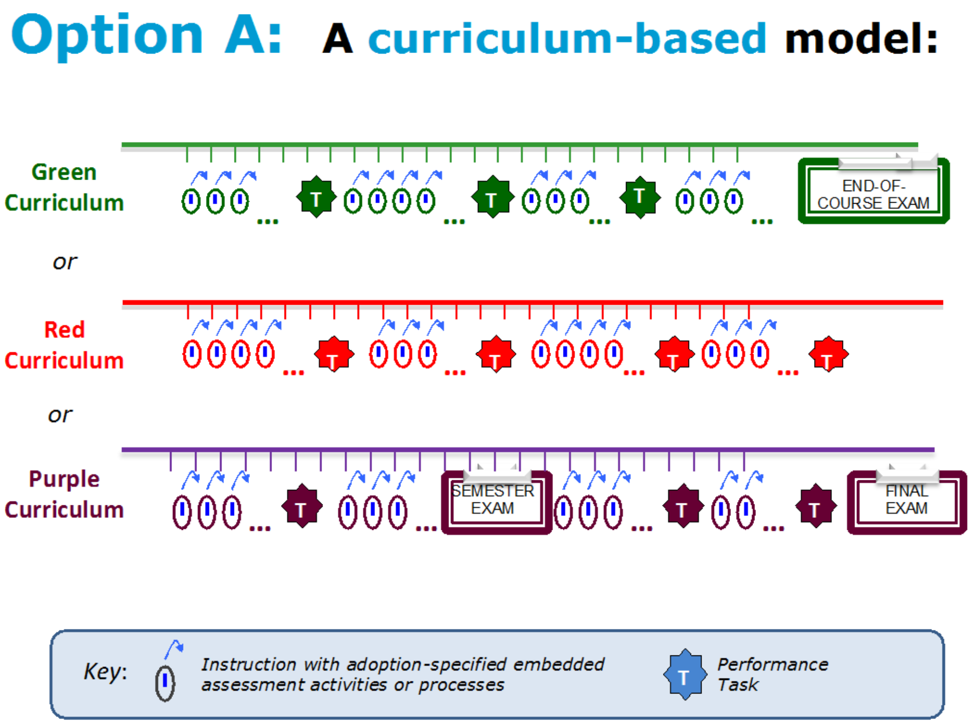


Figure 2. An example of the curriculum-based model.
Illustration Credit: Nancy Doorey.

Any design approach is most fundamentally dependent on the quality of the measures if the NGSS are to be well assessed. A full discussion of measurement quality is outside the scope of this paper, and readers should access the NRC (2014) report for this discussion. The NRC report described in depth for the U.S. science context how high-quality measures must show strong evidence of validity, or evidence that the assessment tasks and the overall system measure the skills that they are intended to measure. The NRC report described facets of validation that include theoretically and empirically evaluating the intended inferences, their use and purposes. Evaluation of alignment, consistency with other measures, fairness, reliability, comparability, user response processes, accessibility, quality control, appropriateness, consequences, and many other factors can enter into validation and use of measures. For any high-quality assessment system to be implemented, the design must describe a strong case for how the validation argument is made, implemented, and evaluated. This will need to be true whether measures are established through adoption processes as described in this section, shared and secure tasks as in the next design example in the Common Tasks Model section, technology-enhanced banks as in the third example in the Traditional With an Inverted Emphasis Model section, or any other of many possible system design approaches.

Of course, as is true in all such matters, states and localities would need in their science and science education communities to have strong, reflective discourse on what they would like to see in their evidence case and their materials. Alignment, for instance in a system model, likely cannot be left entirely to teachers, who unsupported by broader efforts would have insufficient time and lack pooled resources. Policymakers therefore need to think about what they would choose to specify for alignment approaches, as currently can be seen in many U.S. adoption processes, but would be new here, of course, as NGSS is new.

More broadly, they would need to think about what they choose to track in their assessment systems, such as school surveys and course completion rates, and how programs would be evaluated on their success. Such norms are currently in place to some degree in AP and IB systems, but of course this has taken sustained time, energy, and commitment to science education to bring about effectively.

Key to the curriculum-based approach, of course, is that the materials of the assessment systems do a robust job of serving as high-quality measures, including providing a sufficient degree of comparability, and as signifying and instantiating the vision and goals of the NGSS. For this, the observations taken and the interpretations and inferences made would be a key outcome. For instance, with AP and IB science courses, a long tradition has been established of specifying the frameworks for the equivalent of the student model, the task and instrument designs for the equivalent of the task model in the course materials, and the psychometric and other models for assembling and reporting evidence. These would most likely be necessary in the curriculum-based model approach as well.

Also, professional development is extensively involved. For instance in the AP and IB examples and in the Stanford University courses, extensive teacher moderation and teacher professional development are in the use, including for administration, follow-up, feedback, and reporting of assessments. In IB examinations, assessments are often practiced throughout the year and then a work sample set submitted for external moderation by other teachers. A curriculum-based model, or an assessment system that incorporates even a portion of such a model through the adoption process, means building capacity in the science education system in the United States, with all this both entails and promises regarding building an understanding of assessment literacy.

These five issues are concerns in each of the designs—high-quality measures; the need for robust, reflective discourse within states and localities as systems are designed; the need to signify and instantiate the vision and goals of the framework through the design; the call for sustained commitment to assessment literacy; and the need for capacity building and investment in teacher professional development. Subsequent sections will refer back to this discussion for reference on these topics.

Of course, many limitations can be inherent in assessment designs. Limitations are discussed in the Limitations of the Prior Examples section in this paper after all the examples have been introduced.

Common Tasks Model

In the common tasks model, the emphasis is on states and localities working together to develop a library of shared, high-quality assessment resources. The design is called a *common tasks model* because the library would need to include a range of standardized common high-quality *extended* science performance tasks, some shared and some secure.

Extended tasks mean that the observations of student work involve respondents engaged actively and deeply in reflection and often over a period of time. Some tasks might take perhaps more than one sitting, while others might be 15 or 30 minutes long. The key is that students are offered enough time to fuse their knowledge across the NGSS elements—on-demand response cannot be in an instant but takes some time for this type of scientific thinking.

For the common tasks model, some tasks in the bank could be shared transparently with teachers and students, while other tasks are kept secure and used during assessment windows specified by the participating groups, depending on the tasks' system design.

The common tasks would need to bring together the three NGSS dimensions and be well-prepared to align to the framework and standards of the NGSS. Examples of many such tasks can be seen in NRC (2014). These tasks should also be high quality in their psychometric and measurement attributes, such that teachers, students, and schools can have confidence in their use.

Figure 3 shows a multistate shared digital library with common tasks. Some tasks are indicated as shared, meaning they are available for school perusal in the library. Schools might use the tasks at meaningful intervals as shown in Figure 3. The tasks could be transparent to teachers, meaning the content and structure is known to them in advance, if so desired, in the system.

A stream of such resources might be followed at the end of instruction with a secure common task or tasks, also of high quality and aligned to the NGSS. The array of components shown is just one possible high-level combination of tasks. Decisions would need to be made within the system as to the degree of alignment between assessment and instruction, but the model proposed here presumes that main alignment is between each component and the NGSS directly. In other words, tasks are aligned to the NGSS, and the instruction is aligned to the NGSS. Of course, opportunity to learn along the lines of the NGSS goals and objectives must take place for meaningful and useful assessment of outcomes. This is a given in this model.

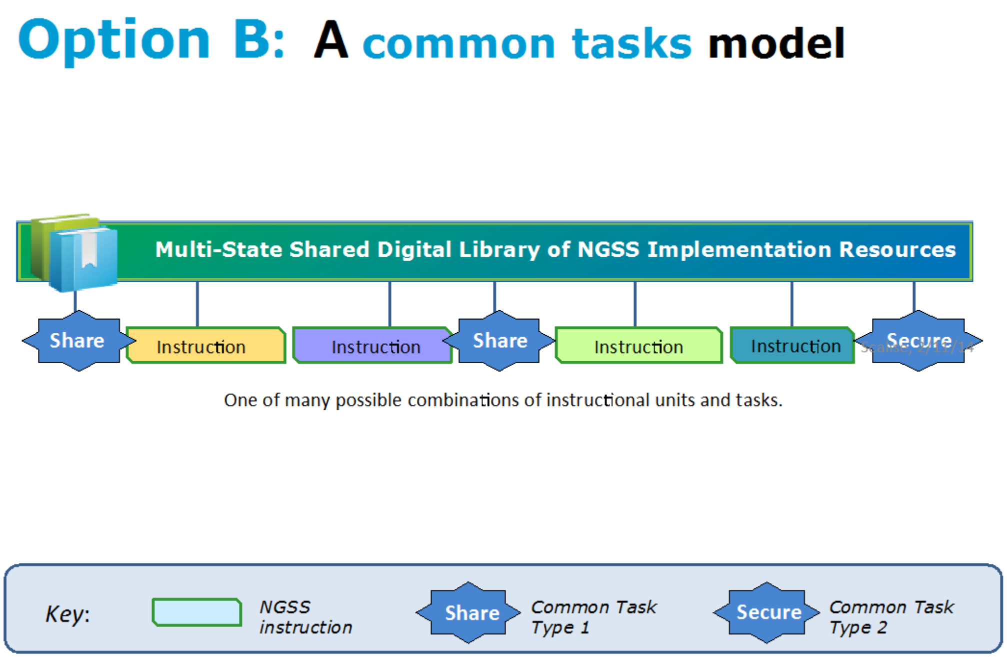

Option B: A common tasks model
Multi-State Shared Digital Library of NGSS Implementation Resources
Share
Instruction
Instruction
Share
Instruction
Instruction
Secure
One of many possible combinations of instructional units and tasks.
Key:
NGSS instruction
Share
Common Task Type 1
Secure
Common Task Type 2

Figure 3. An example of the common tasks model.
Illustration Credit: Nancy Doorey.

Tasks familiar to states from other assessments that fit a common task model range from shorter tasks, such as the State of Oregon work samples, to fully self-contained replacement units, as have been used in Delaware. The longer approach sometimes includes a menu of instructional units—as much as 3–4 weeks worth—that are provided to schools, perhaps with kits, materials, and other resources, on which the common assessments are then based.

Common tasks, whether shared or secure, as described in the Curriculum-Based Model section of this paper can be teacher moderated or technology enhanced for scoring. Elements of choice and assignment can be established within the system design parameters. These elements could include a selection for a range of tasks or a selection of delivery windows and other aspects of administration and logistics.

State monitoring in this model would likely need to consist of not only the secure tasks and questions associated with them, but also the opportunity to learn information sought from schools such as input from school and district surveys. Examples of these types of input surveys that would be valuable in making policy decisions and improvements in student-learning outcomes as described by the student model include the NAEP science state report cards, which might be an informational element in such a system, as well as other NCES educational surveys that are available to supply evidence on characteristics of student learning that might be helpful to understand. Internationally, the OECD PISA assessments and their background questionnaires have been used by some countries to enhance and inform their educational systems.

As described in more detail previously, a number of key commitments are necessary including robust, reflective discourse within states and localities as systems are designed; the need to signify and instantiate the vision and goals of the framework through the design; the call for sustained commitment to assessment literacy; and the need for capacity building and investment in teacher professional development.

Traditional With an Inverted Emphasis Model

The traditional with an inverted emphasis model requires some explanation. What is traditional in assessment, after all, and what is meant by inverted?

Traditional in this case is not very old, but rather is meant to describe the types of assessment systems developed over about the last 5 years from state consortia working together in the United States, in other subject matter areas that did not typically include science assessment. Six Race to the Top (RttT) consortia were mentioned previously in the paper. Two of these consortia—PARCC and Smarter Balanced—developed systems that were outlined and described at the Invitational Research Symposium on Science Assessment, providing much food for thought in the science community.

The efforts of the RttT consortia represent many innovations and some important advances in assessment (Scalise, 2013b), especially in mathematics and reading. The systems also have faced many challenges developing some common assessments in the United States. From the point of view of science assessment, there are lessons to be learned. Here I use the term *traditional* because an expectation is that systems such as these consortia represent are fast becoming a norm in many, although not all, regions, and most states now have some understanding of what is incorporated in the six approaches, whether a state is participating fully or not.

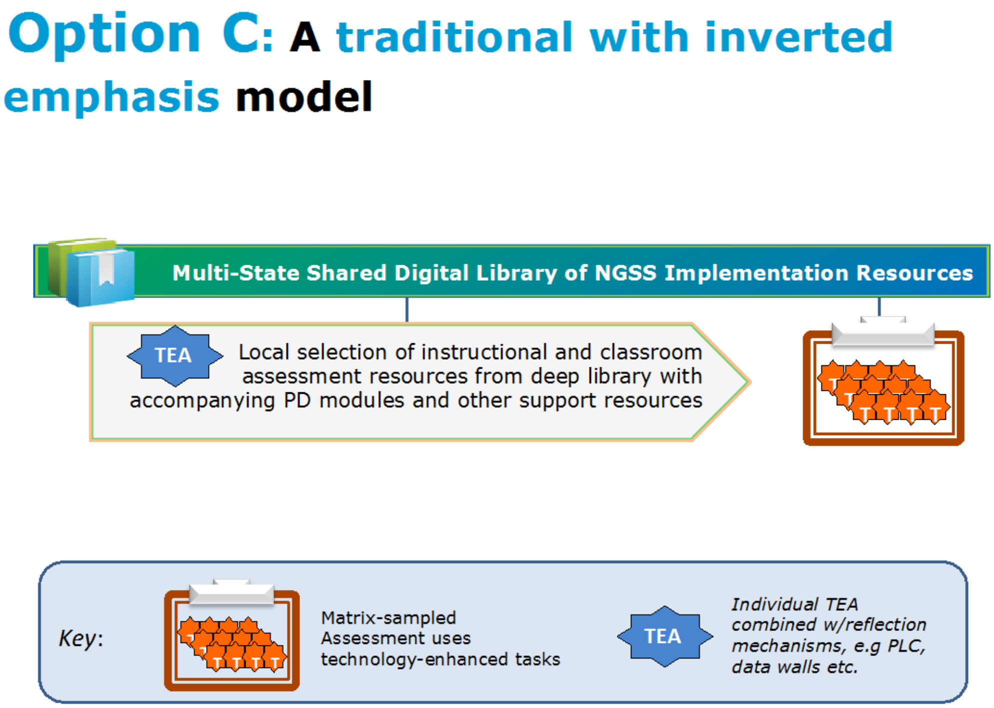
However, I use *inverted* here *to* indicate that the focus of the model is specifically on a need for science assessment that is different from how the math and reading developments have progressed. Under the RttT Assessment Program (U.S. Department of Education, 2010), grantees were required to place top priority on the development of summative assessments in English language arts and mathematics for students in Grades 3 through 8 and high schools that would produce individual student results starting in 2014–2015 to meet all federal accountability provisions under the Elementary and Secondary Education Act. These systems were also to produce data to address four additional needs: (a) determinations of school effectiveness; (b) determinations of individual principal and teacher effectiveness for purposes of evaluation; (c) determinations of principal and teacher professional development and support needs; and (d) teaching, learning, and program improvement. These systems, then, were to prioritize accountability purposes over instructional purposes.

The intent, then, in the use of *inverted* in the name of this system design is to signal that the priority here is placed on producing assessment information for instruction in the classroom, with an additional but not superseding priority placed on production of summative data for monitoring purposes. Science in the United States has different needs. As so eloquently stated by a classroom teacher serving on the NRC committee that generated the NRC NGSS report (NRC, 2014), she, like so many science educators, has struggled throughout her career with the capacity in the U.S. educational system to provide the opportunity to learn science for students. Science has few instructional minutes in the United States, even fewer devoted exclusively to science instruction—so much so that some schools in the elementary grades have focused on an avenue of reading and writing literacy to incorporate additional science minutes. In some states, few students take many credits of science during high school. Yet whatever the student's age, PISA studies have found strong links between learning time and educational outcomes.

Additionally, science educators in the United States, especially in recent years, often have had few current materials available for them to use with students. The burgeoning need to add resources for basic instruction in other areas, along with shrinking dollars for education, has meant hard-pressed schools have little remaining for subject matter areas such as science. For instance, the NRC committee teacher described how her classroom had not had a new adoption of science materials for 17 years.

Teachers engaging in science activities many times have proven resourceful in what they can do. Yet if a resource such as enhanced assessment information is being considered for science education, the NRC (2014) report calls for inverting the focus due to the capacity needs in science education and ensuring the information is supplied in sufficiently large measure to the classroom. This is what is meant by an inverted focus. Considered in this model is the design of systems that might leverage the infrastructure being created by the RttT consortia, but also might direct an inversion of resources—more of the state discourse conversations, consortia reflection, system building efforts, and ultimately infusions of assessment information to support the opportunity to learn directly. This may be required in science education, where capacity building is an issue.

So in the model shown in Figure 4, investments such as the RttT infrastructure are leveraged to create a variety of robust NGSS-aligned technology-enhanced assessments (TEA) developed to serve classroom needs and provide critical learning evidence for all students. State monitoring takes place through a relatively small group from these TEA standardized tasks, which are matrix-sampled to provide the desired level of information across each locality, state, or region participating, depending on the intended design of the system.



Option C: A traditional with inverted emphasis model

Multi-State Shared Digital Library of NGSS Implementation Resources

TEA Local selection of instructional and classroom assessment resources from deep library with accompanying PD modules and other support resources

Key: Matrix-sampled Assessment uses technology-enhanced tasks

TEA *Individual TEA combined w/reflection mechanisms, e.g PLC, data walls etc.*

Figure 4. An example of a traditional with inverted emphasis model.

Illustration Credit: Nancy Doorey.

Additionally, at the school level, portfolios of individual TEAs could be combined with reflection mechanisms for professional development, such as in a professional learning community (PLC) approach. Discussing PLCs in depth and their use in assessment is outside of the scope of this paper, but is mentioned briefly here as a possible component of the model.

Key to the inverted approach may be the ability of states and/or localities to work together on matrix-sampling designs. In the traditional but inverted emphasis model, all students are called to participate in high-quality assessments, thereby having high-quality actionable information relevant to instruction made available to them and their teachers to advance the building of students' scientific knowledge, skills, and abilities. For additional monitoring, a small number of mostly performance tasks such as hands-on or remote hybrid labs could be exciting tasks for schools from which to pool data. For those activities that are matrix-sampled, any given student might engage in only one or a few such experiences over time. The information flows together with data from other students for an aggregated monitoring picture. Together the evidence collected across students would then need to be used in ways effective for monitoring of the system, as described in the NRC (2014) report.

As described previously in more detail in the first and second model options, a number of key commitments are necessary for this model, including robust, reflective discourse within states and localities as systems are designed; the need to signify and instantiate the vision and goals of the framework through the design; the call for sustained commitment to assessment literacy; and the need for capacity building and investment in teacher professional development.

Limitations of the Prior Examples

It should be pointed out that the three models presented so far are by no means intended to be an exhaustive set of possible approaches. Rather, they simply illustrate ways of thinking through some evidence-centered design ideas. The models were specifically selected to provide a strong contrast among the set, so that it is clear how they differ.

For that reason, the models tend to be extreme in the design space. A more moderated version, that include gradual transitions or blending elements of the designs might be desirable for each model. The first model, for instance, could benefit from a graduate transition. This model requires arguably greater interaction for many state and local assessment personnel with their colleagues, so may take some transition time to unpack the new standards. The second and third models require the development of likely substantial shared resources. Validation and use of new shared and secure tasks, or technology-enhanced banks, could be gradual, or might be blended with some currently existing state or local resources as appropriate.

The NRC report (2014) argued that costs will be incurred if the NGSS are to be fully assessed. The NRC report recommended that priorities should begin with what is both necessary and possible in the short term *but at the same time* establish long-term goals to implementation of a fully integrated and coherent system of curriculum, instruction, and assessment. As the NRC report described, assessment system approaches that the committee recommends may differ in some important respects from current practice. Time will also be required to adopt the instructional programs needed for students to learn science in the way envisioned in the framework and the NGSS.

The models presented here as examples also are quite skeletal, or not fully described in many important aspects. For brevity of this paper, the basic outlines of the models are sketched; much more complete details would be necessary to implement any actual system.

A Nonexample, for Contrast

I include a nonexample here because sometimes an example of what might *not* be an NGSS-aligned system is unclear to some stakeholders involved in the discussion. Educators should remember that not everyone interested in learning outcomes for U.S. students has as much time to spend in the classroom, so often more examples and their implications can be helpful.

The nonexample here is intended to show some direct conflicts with the NGSS goals. The nonexample is a 40-minute once per year examination with 40 separate questions, given once in each grade from kindergarten through eighth grade and twice in high school, with questions aligned to the grade level at each grade.

The extremes of the nonexample will be pushed further, as an illustration, and it will be assumed the yearly examination is the sole source of information—the only component of a system of science assessment—and is intended to be used extensively in program evaluation of various types, including the fidelity and efficacy of NGSS science instruction in the classroom. The nonexample is also burdened with the assumption that the results of the examination are expected to be useful to understand other student characteristics in science, such as students' developing attitudes, dispositions, and communication skills in the sciences, and their opportunity to have studied in the NGSS goals and objectives.

One argument here for this situation as a nonexample draws on conclusions of the NRC NGSS report (NRC, 2014). Forty separate, unrelated questions in 40 minutes allow students a very brief time to think in reflective ways and draw on their NGSS-related knowledge, skills, and abilities. According to the NRC (2014) report, students would have too limited opportunity in 40 minutes on 40 separate questions to bring to bear the practices and crosscutting concepts to exemplify the core knowledge on which they have worked throughout the year and to show what they know and can do. It would not be an adequate assessment; thus student understanding would not be adequately measured and reported.

So time is a big issue here. Context is also an issue. Students being assessed in NGSS style need to be able to pose questions, engage in use of theories and models, and examine evidence. This kind of scientific thinking cannot be done in a dearth of context. Also, information and tools are needed, at least to some extent, whether for instruction or assessment of NGSS. Presumably, so many different, independent questions in such a short time could not incorporate much context for any given question.

Due to such issues as these, results might be statistically reliable—meaning, for instance, another set of similar questions might produce similar results—but not valid. These questions' appropriateness as adequate measures of NGSS, because they do not tap NGSS skills, would be greatly in question. Thus the ability to serve the informational needs of stakeholders, much less at multiple levels of the system, would be jeopardized. Given the limitations, these questions would not be accurate measures of the NGSS.

Furthermore, the role of such questions in signifying the goals of NGSS would be in conflict. Teachers and students seeing the types of assessment questions being asked by this nonexample could also be expected to treat these candidate questions as role models or examples of ideal types for the NGSS. Otherwise, if they were not the best questions, teachers ask, why would the states and localities provide them?

As discussed previously, assessments have a strong role in signifying what should be taught and worked on in the classroom. Proxy measures that do not exemplify the real thing can be easily misinterpreted—teachers often do not understand them as having some predictive validity. In some cases, proxy assessments may deliver destructive impacts through consequential validity involved in their use and in the response processes of the educational system as a consequence of their use.

To mention one more concern for the nonexample, it is not a systems view. The single yearly examination of this type, again admittedly an extreme example for contrast, would provide only one type of evidence to the educational system. Arguments in support of the need for a coherent system of information are explored in depth in a prior section. So suffice to say that few, if any, of the system considerations described there are met by the single, short, yearly assessment described here in this particular format.

In terms of achieving the goals of the NGSS, the nonexample misses the match with the student model in some areas. Although such an approach as a single short yearly test of this format may be perceived as practical and simple to administer, states and localities still should ask what would it accomplish to achieve the student model for any particular model or approach.

Of course, moving to other solutions if necessary will take time and effort, as described in the Limitations of the Prior Examples section in this paper a. The time and effort needed suggests measured approaches, gradual efforts, and incremental change may suit state efforts as they consider system designs to fully assess the NGSS.

Conclusion

In summary, this paper provides some high-level examples of designs for comprehensive science assessment systems that could potentially align with the NGSS. The designs are specifically sampled to be extreme cases in order to depict the differences among them better in terms of their approaches and how they satisfy their purposes.

High-level components of a curriculum-based model, a common tasks model, and a traditional with an inverted emphasis model are illustrated. It is important to remember, however, that in any given system design, elements could be shared between models, moved from one model to another, or be replaced with other ideas, and of course many other models could be created. Regardless, what needs to remain is a coherent system, based on a student model that shows how the assessment information will make a difference in student learning outcomes. Clear connections between the elements and their use should be in place.

As described in the NRC NGSS report (NRC, 2014), aspects such as practicality under a given set of circumstances, prior infrastructure for assessment available to leverage, and of course the inclinations of those involved also enter into the design discussion. These aspects may take the form of prioritizing what is considered for use. However, by approaching the design from a principled basis built on a theory of action for students, the connection between logistics and utility is not lost. The student model, or goal of improving learning outcomes through the assessment information, is not forgotten. In other words, information should not simply be collected for information sake. Practical and preference aspects are of course an imperative, but they should not be the only or sole basis on which system designs are built.

More important than the specific designs shown here is the concept of how to think about designing a system in general. What are the essential questions, and how might they be framed? How, in essence, might states or localities think about entering into the process of thinking about assessing in the NGSS context? Here, the approach taken is to borrow from the concepts of evidence-centered design, which has been applied extensively in the U.S. context to design observations or tasks, and more recently to design instruments or test forms. These ideas are extended here to apply the concepts to the overall system design itself. Figures 2–4 specifically frame the student model, task model, and evidence model for potential systems views.

The intention of this paper is to describe some overall ways that system design might be approached for NGSS, drawing on the NRC assessment triangle, its subsequent reports, and concepts of evidence-centered design practices. Then secondly, this paper offers three possible high level designs of comprehensive science assessment systems, ranging from perhaps least familiar in the U.S. context to most familiar.

The recommendation here is for states and localities to think through, with robust and reflective discourse, how to make a coherent design argument as a useful and essential part of developing a science assessment system that would serve learners well. Basing decisions for an assessment system on enhancing learning outcomes only makes sense. But this means having high-quality information available to meet the theory-of-action goals of the system. The NGSS explain how important the use of evidence is for students in science education. This paper describes some ways to think about evidence—connections between hypotheses, collections of evidence, and conclusions—in the assessment of NGSS outcomes.

Author Note

This material is based upon work supported by the U.S. National Science Foundation under Grant Nos. 0737056 (CCLI) and 0722014 (REESE), and by Cisco, Intel, and Microsoft through ATC21S, atc21s.org.

References

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., . . . Steele, C. M. (2013). *Criteria for high-quality assessment.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

DeBarger, A. H., Penuel, W. R., & Harris, C. J. (2013). *Designing NGSS assessments to evaluate the efficacy of curriculum interventions.* Paper presented at the Invitational Research Symposium on Science, Washington, DC. Retrieved from the K-12 Center at ETS website: http://www.k12center.org/rsc/pdf/debarger-penuel-harris.pdf

Embretson, S., & Reise, S. P. (2000). Item response theory as model-based measurement. *Item response theory for psychologists* (pp. 40–61). Mahwah, NJ: Lawrence Erlbaum.

Haertel, G. D., Cheng, B. H., Cameto, R., Fujii, R., Sanford, C., Rutstein, D., & Morrison, K. (2012). *Design and development of technology-enhanced assessment tasks: Integrating evidence-centered design and universal design for learning frameworks to assess hard to measure science constructs and increase student accessibility*. Paper presented at the Invitational Research Symposium on Technology-Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performanceassessments. *Educational Researcher, 23*, 13–23.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (CRESST Technical Paper Series). Los Angeles, CA: CRESST.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Research Papers in Education, 25*(3), 253–270.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2005). *Systems for state science assessment*. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*, 320–323.

Quinn, H., Schweingruber, H., & Keller, T. (Eds.). (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

Scalise, K. (2012). *Using technology to assess hard-to-measure constructs in the CCSS and to expand accessibility*. Paper presented at the Invitational Research Symposium on Technology-Enhanced Assessments, Washington, DC. Retrieved from the K-12 Center at ETS website: http://www.k12center.org/events/research_meetings/tea.html

Scalise, K. (2013a). NGSS assessment system designs: System design options. Paper presented at the Invitational Research Symposium on Science Assessment, Washington, DC.

Scalise, K. (2013b). *Planning for transition: Evolution of next-generation assessment system designs to support ongoing improvements in measuring complex skills and constructs*. Presentation at the CCSSO National Conference on Student Assessment, National Harbor, MD. Retrieved from https://ccsso.confex.com/ccsso/2013/webprogram/Session3431.html

Scalise, K. (2014, April). *mIRT-bayes as hybrid measurement model for technology-enhanced assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

U.S. Department of Education. (2010). *Overview information*. Race to the Top Fund Assessment Program. Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010. http://www2.ed.gov/legislation/FedRegister/announcements/2010-2/040910e.html

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716–730.

The Center for K–12 Assessment & Performance Management at ETS creates timely events where conversations regarding new assessment challenges can take place, and publishes and disseminates the best thinking and research on the range of measurement issues facing national, state and local decision makers.